Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.35240

A Study on Prompt Engineering Strategies and Context-Enhanced Translation for Video Game Localization

Yuanze Zou

Department of English (Language Data Science), School of Foreign Studies, Xi'an Jiaotong University, Xi'an, China
prejudicezyz@foxmail.com

Abstract. Video game localization often depends on contextual understanding rather than sentence-level translation alone. This problem is especially evident in fragmented narrative games such as Elden Ring, where information about characters, places, factions, terminology, and world-building is scattered across item descriptions, equipment texts, skill explanations, and other isolated entries. When these entries are translated without sufficient background information, large language models may produce inconsistent terms, weakened narrative references, inaccurate interpretation of polysemous expressions, or a style that does not match the original atmosphere. To address this problem, this study proposes a context-enhanced prompt-based translation framework for game localization. The framework builds a local semantic knowledge base, retrieves relevant background information through contextual embeddings and cosine similarity, and inserts the retrieved context into structured translation prompts. Using Elden Ring item descriptions as the test corpus, this study compares Doubao, Gemini, and GPT-5.5 under four settings: baseline translation, prompt engineering, context enhancement, and the full method. The evaluation uses BERTScore F1, character-level BLEU-4, and the Global Terminology Enrichment Score (GTES). The results show that context retrieval is more effective than prompt engineering alone in improving domain terminology recall, and the full method achieves the strongest overall performance. This suggests that external knowledge retrieval can help reduce contextual loss in fragmented game text translation.

Keywords: video game localization, large language models, prompt engineering, context enhancement, retrieval-augmented generation

1. Introduction

As video games continue to expand their influence in the global digital entertainment industry, video game localization is no longer merely a matter of linguistic transfer. It has become a cross-cultural practice involving narrative reconstruction, cultural adaptation, terminology management, and user experience. In role-playing, open-world, and narrative-driven games, texts not only provide functional instructions but also carry world-building information, imply character relationships, and shape stylistic atmosphere. Therefore, the quality of game text translation directly affects players' understanding of narrative logic, cultural imagery, and aesthetic style.

DOI: 10.54254/2755-2721/2026.35240

Fragmented narrative games such as *Elden Ring* pose particular challenges to traditional localization workflows. Their stories are not presented through continuous linear narration. Instead, world-building settings, mythological structures, character relationships, and core terminology are distributed across item descriptions, equipment texts, spell descriptions, place names, and character titles. A single text entry is often only one node in a broader narrative network, and its meaning needs to be supplemented by other entries. However, in actual localization workflows, game texts are often stored and translated as isolated strings, making it difficult for translators or machine translation systems to obtain sufficient contextual information.

Large language models have created new possibilities for video game localization, but their general translation ability does not fully solve context-dependent translation problems. Existing research shows that ChatGPT performs relatively well in high-resource European languages, but remains weaker in low-resource or distant languages, as well as in domain-specific and noisy texts such as medical abstracts and Reddit comments [1]. This limitation is relevant to game localization, where translation difficulties often come not only from language transfer, but also from terminology systems, fictional world-building, and fragmented narrative structures.

Prompt engineering has been regarded as one way to improve the controllability of large language model translation. Translation-related information such as purpose, target audience, locale, register, and style guide can influence model output [2]. However, such prompts usually provide macro-level task instructions and may fail to capture the intertextual relations and fine-grained background knowledge hidden in fragmented game texts. Further research also suggests that prompt strategies such as translation brief, translator persona, and author persona do not necessarily produce statistically significant improvements [3]. Therefore, context support for large language model translation should not remain limited to role setting or general style instructions. It requires more precise, retrievable, and text-specific knowledge injection.

Recent work on human-like translation strategies also suggests the value of pre-translation knowledge preparation. The MAPS framework requires the model to extract keywords, topics, and related examples before generating and selecting translations [4]. However, this type of auxiliary knowledge is mainly generated by the model itself and may still introduce noise or hallucination. For game texts such as *Elden Ring*, where terminology systems and world-building relations are relatively stable, a more suitable approach is to externalize the knowledge source by using a local semantic knowledge base to provide traceable contextual information.

Based on this research gap, this study proposes a context-enhanced prompt-based translation framework for fragmented narrative game texts. The framework extracts contextual embedding vectors from the source text, retrieves relevant background information from a local semantic knowledge base using cosine similarity, and dynamically injects the retrieved context into structured prompts. Studies on contextual embeddings show that models such as BERT can encode sentence-level linguistic information and context-related features, which supports the use of semantic vectors for context retrieval [5].

This study uses item descriptions from *Elden Ring* as the research corpus and selects Doubao, Gemini, and GPT-5.5 for cross-model experiments. Four experimental conditions are established: the baseline group, the prompt engineering group, the context enhancement group, and the full method group. To evaluate translation quality, this study uses BERTScore F1, character-level BLEU-4, and the proposed Global Terminology Enrichment Score (GTES), which measures the recall of key terms, world-building concepts, and game attribute terms.

The main contributions of this study are as follows. First, it constructs a dynamic prompt-based translation framework using contextual embeddings and semantic retrieval to address contextual

DOI: 10.54254/2755-2721/2026.35240

isolation in fragmented game texts. Second, it builds a local semantic knowledge base for *Elden Ring* and uses cosine similarity to match source texts with relevant contextual information. Third, it conducts a cross-model, four-group ablation experiment to examine the effects of prompt engineering and context enhancement. Fourth, it proposes GTES as a domain knowledge evaluation metric to complement BERTScore and BLEU-4 in assessing terminology and world-building consistency.

2. Literature review

2.1. Video game localization and video game machine translation

Game localization cannot be reduced to replacing words in one language with words in another. In an actual production workflow, localization also involves technical adaptation, cultural adjustment, testing, release coordination, and the control of player experience [6]. O'Hagan and Mangiron also place game localization within the wider global game industry, where language, gameplay, cultural expectations, and user experience are closely connected [7]. For this reason, a localized game text needs to be accurate at the sentence level, but it also needs to preserve narrative coherence, gameplay readability, and cultural acceptability.

Another important feature of game localization is that translators often work creatively under restrictions. Text length, interface space, platform requirements, variables, and game mechanics can all limit translation choices [8]. This is one reason why video game translation has been discussed as a specialized field rather than a simple branch of literary translation, audiovisual translation, or software localization [9]. Game texts also have multimodal features: written language, images, sound, interface elements, and player interaction may all influence how a line should be understood and translated [10].

Machine translation has gradually entered this field, but its role remains limited and task-dependent. Research on English–French game translation shows that a customized system trained on in-domain game data can outperform general public translation systems [11]. This finding suggests that game-domain data are useful, but it does not mean that machine translation can replace human localization work. A more realistic use is to treat it as an auxiliary tool for handling large quantities of repetitive or semi-repetitive text [11]. One persistent difficulty is that game texts are often stored in spreadsheets or string files. Information about speaker, scene, tone, medium, and interactive function may be missing. This is close to the "isolated string" problem examined in this study: without context, even a fluent translation may fail to preserve the intended term, relation, or narrative implication.

2.2. Retrieval-augmented generation and domain knowledge injection

Retrieval-Augmented Generation offers one possible way to address the knowledge gap of large language models. In the original RAG framework, the model does not rely only on knowledge stored in its parameters. It retrieves relevant information from an external document collection and uses that information during generation [12]. The framework combines parametric memory from a pre-trained sequence-to-sequence model with non-parametric memory from a dense-vector-indexed document collection, which makes the knowledge source more flexible and easier to update [12].

This design is useful for tasks where factual or domain-specific information matters. A language model may generate fluent text, but it may not always access the right knowledge, especially when the required information is specific, newly updated, or scattered across documents. External retrieval

DOI: 10.54254/2755-2721/2026.35240

can therefore provide a more traceable source of evidence and reduce unsupported generation. Recent reviews also note that RAG can improve the relevance and factual support of large language model outputs in domain-specific tasks [13].

Most RAG studies focus on fields such as question answering, fact verification, medicine, finance, software development, and other knowledge-intensive scenarios [13]. Video game localization has received less attention in this line of research. However, fragmented narrative games such as *Elden Ring* also contain knowledge-intensive features: proper nouns are used repeatedly, character relations are implied across entries, and world-building information is distributed rather than explained in one place. For this reason, this study applies RAG-style context retrieval to game localization. By building a local semantic knowledge base and retrieving relevant background information through contextual embeddings and cosine similarity, the framework aims to provide missing context for isolated game text entries.

2.3. Large language models, prompt engineering, and translation customization

Large language models have changed the way machine translation tasks are formulated. Instead of using a fixed translation system for a fixed language pair, users can give the model instructions about target language, style, audience, and translation purpose. This makes LLM-based translation more flexible than many traditional neural machine translation systems, especially when the task requires style control or additional explanation.

At the same time, this flexibility does not mean that LLM translation is equally reliable in all situations. Research on ChatGPT shows that its performance differs across language pairs and text types. It performs better in some high-resource European language pairs, but its results are less stable in low-resource languages, distant language pairs, domain-specific texts, and noisy materials [1]. Pivot prompting can partly improve translation between distant languages by introducing a high-resource intermediate language [1]. However, this strategy mainly deals with language distance. It does not directly solve the type of problem discussed in this study, where the difficulty comes from fictional terminology, hidden narrative relations, and missing contextual information.

Prompt engineering is another way to guide LLM translation. Previous research shows that adding information such as translation purpose, target audience, register, and style guide may affect ChatGPT's output and make the result closer to practical translation requirements [2]. For game localization, however, this kind of prompt is still too general. A prompt can tell the model to use a dark fantasy style, but it cannot by itself identify which character, faction, location, or world-building concept is implied in a short item description.

This limitation is also reflected in studies on translation brief and persona prompts. A comparison of translation brief, translator persona, and author persona prompts shows that these prompt types do not necessarily lead to statistically significant improvement [3]. This does not mean that prompt design is useless, but it suggests that macro-level instructions alone are not enough for highly context-dependent translation. For this reason, the present study uses prompt engineering together with semantic retrieval, so that the prompt contains not only general translation instructions, but also text-specific background information retrieved from a local knowledge base.

2.4. Knowledge-guided translation, semantic retrieval, and evaluation

Traditional machine translation usually treats translation as a direct mapping from the source text to the target text, while professional translators often conduct text comprehension, terminology confirmation, background checking, and stylistic judgment before translation. The MAPS

DOI: 10.54254/2755-2721/2026.35240

framework simulates this process by requiring large language models to generate translation-related knowledge, including keywords, topics, and relevant examples, before producing and selecting candidate translations [4]. Experimental results show that MAPS can improve translation performance and reduce problems such as mistranslation, awkward style, untranslated content, and omissions [4]. This suggests that pre-translation knowledge preparation is useful for improving LLM translation. However, MAPS mainly relies on model-generated auxiliary knowledge, which may introduce noise or hallucination, and its serial processing also increases inference time [4].

This study therefore adopts an external knowledge base retrieval approach. Instead of asking the model to generate background knowledge by itself, the proposed method retrieves relevant information from a local semantic knowledge base. This makes the knowledge source more closely related to specific game texts, more stable in terminology and world-building information, and easier to trace and maintain. In this sense, the proposed method can be regarded as an external knowledge base-enhanced version of human-like translation strategies.

The technical basis of this retrieval process is contextual embedding and cosine similarity. Contextual embedding models can generate dynamic representations according to the context in which words appear, making them more suitable than static word embeddings for handling polysemy, metaphorical expressions, and semantic dependencies. Existing research comparing BERT and Word2vec shows that BERT can encode information involving the entire input sequence, especially in tasks related to syntactic features [5]. This supports the use of contextual embeddings to represent source texts and retrieve relevant contexts in this study.

In terms of evaluation, traditional automatic metrics such as BLEU mainly measure surface matching, while neural metrics such as BERTScore and COMET focus more on semantic similarity. However, automatic metrics may not fully capture fine-grained differences caused by specific prompts or translation strategies, and BLEU and COMET were not originally designed to evaluate subtle prompt effects within the same system [3]. Therefore, this study uses BERTScore and BLEU-4 together with the proposed Global Terminology Enrichment Score (GTES), which measures the recall of core terminology, world-building concepts, and game attribute terms. For terminology-intensive and narratively fragmented texts such as those in *Elden Ring*, GTES helps compensate for the limitations of general automatic metrics in evaluating domain knowledge preservation.

3. Methodology

3.1. Research objectives

The main purpose of this study is to test whether external contextual information can improve large language model translation in video game localization. The problem is not only whether the generated Chinese text is fluent, but whether the model can keep the style, terminology, and background relations of fragmented game texts. Based on this concern, the experiment is organized around three questions: whether structured prompts can help control style and terminology; whether vector-based retrieval can provide useful background information for isolated text entries; and whether the two methods work better when they are combined.

To answer these questions, this study uses item description texts from *Elden Ring* as the experimental corpus. A local semantic knowledge base is built to provide background information, and a terminology glossary is prepared for domain-term evaluation. Four experimental settings are used: baseline translation, prompt engineering, context enhancement, and the full method. By comparing these settings, the experiment separates the effect of prompt instructions from the effect of retrieved context, and then examines whether the full framework brings additional improvement.

DOI: 10.54254/2755-2721/2026.35240

3.2. Overall technical framework

This study designs a vertical-domain translation framework named KT-LLM, which refers to Knowledge Base–Test Corpus–Large Language Model. The framework does not fine-tune the translation models. Instead, it changes the information provided to the model before generation. The basic idea is to connect each source text with relevant entries in a local knowledge base, and then place the retrieved information into the translation prompt.

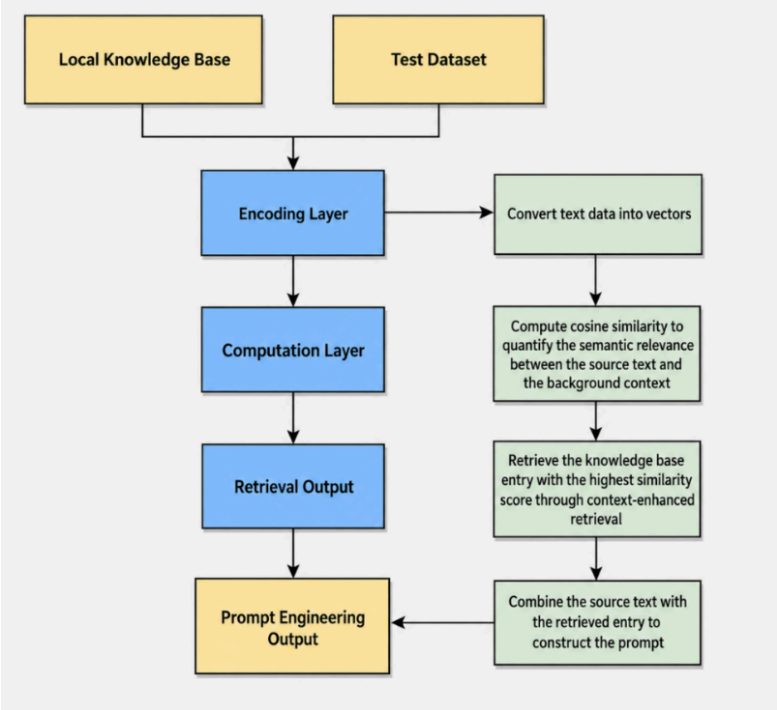

Figure 1. Network structure of KT-LLM

Finally, the large language model generates the target translation based on the enhanced prompt. The framework contains two main parts. The first part is context retrieval. Source texts and knowledge base entries are encoded with the MiniLM model, and cosine similarity is used to find the most relevant background entry for each source text. The second part is prompt construction. The retrieved context is combined with role instructions, style requirements, terminology constraints, few-shot examples, and the final translation task. The model then generates the target translation based on this enhanced prompt.

In the actual workflow, the system first receives an English source entry from the test corpus. The source entry and the knowledge base texts are then converted into semantic vectors. After similarity calculation, the most relevant knowledge base entry is selected as contextual support. This retrieved information is inserted into the prompt template together with the translation requirements. Finally, the large language model produces the Chinese translation. In this process, the knowledge base serves only as an external source of context and does not participate in model training.

DOI: 10.54254/2755-2721/2026.35240

3.3. Module design

3.3.1. Prompt engineering module

In this study, the prompt engineering module is mainly responsible for giving the model a clearer translation frame before it produces the target text. The prompt template contains five parts: translator role, style requirements, terminology requirements, few-shot examples, and the source text to be translated. The role setting asks the model to work as a translator familiar with video game localization and dark fantasy narratives. The style instruction emphasizes a concise, archaic, and solemn tone, which is closer to the textual atmosphere of Soulslike games. The terminology instruction reminds the model to keep character names, place names, proper nouns, and game mechanics terms consistent. A small number of examples are also included to show the expected wording and rhythm of the translation.

This module is not expected to provide new background knowledge. Its main function is to reduce uncontrolled stylistic shifts, casual wording, and unnecessary replacement of established terms.

3.3.2. Context enhancement module

The context enhancement module provides the background information that is missing from isolated game strings. In the experiment, both the source text and the entries in the local knowledge base are encoded with the MiniLM model. After vectorization, cosine similarity is calculated between the source text and each knowledge base entry. The entry with the highest similarity score is selected and added to the translation prompt as external context.

This design is used because many *Elden Ring* texts are short but highly dependent on other entries. A static prompt can tell the model how to translate, but it cannot know which character, faction, place, or concept is relevant to a specific sentence. By retrieving context separately for each source text, this module gives the model more specific background support before generation.

3.3.3. Experimental group design

The experiment uses four groups to separate the effects of the two modules. Group A is the baseline group, where the model receives only the source text. Group B adds the prompt engineering module but does not include retrieved context. Group C adds retrieved background information but does not use the full structured prompt. Group D combines both prompt engineering and context enhancement. This design makes it possible to compare the independent effect of each module and the combined effect of the full method.

4. Experiments

4.1. Experimental environment and parameter settings

The experiment uses *Elden Ring* texts as the test object and selects three large language models—Doubao, Gemini, and GPT-5.5—as translation models. The models are not fine-tuned; instead, their outputs are guided through prompt design and external knowledge retrieval. In the vector retrieval stage, this study adopts paraphrase-multilingual-MiniLM-L12-v2 as the embedding model, with an embedding dimension of 384. Source texts and local knowledge base entries are converted into semantic vectors, and cosine similarity is used to measure their semantic relevance. The retrieval

DOI: 10.54254/2755-2721/2026.35240

strategy is Top-1 retrieval, with no additional similarity threshold. The knowledge base field is Context_text, the source text field is eng_text, and the reference translation field is human_reference.

In the translation generation stage, all groups use the same test texts but different input settings: Group A uses only the source text, Group B adds structured prompts, Group C adds retrieved background information, and Group D combines both prompt engineering and context enhancement. In the evaluation stage, this study uses BERTScore F1, character-level BLEU-4, terminology precision, GTES, and terminology F1. BERTScore is calculated with lang="zh" and idf=False; BLEU-4 is calculated at the character level with equal weights for 1-gram to 4-gram, namely 0.25 each, and SmoothingFunction().method1 is used for smoothing. Terminology-related metrics are based on a manually constructed domain terminology glossary.

4.2. Dataset and data processing

The experimental data are derived from *Elden Ring* game texts and a related local semantic knowledge base. The game has strong fragmented narrative features: world-building information is distributed across item descriptions, equipment texts, character titles, place names, faction names, and attribute descriptions rather than appearing in continuous passages. As a result, individual text entries are highly context-dependent, and translating them as isolated strings may lead to terminology inconsistency, semantic drift, and stylistic generalization.

The dataset consists of three resources: the test corpus, the local semantic knowledge base, and the domain terminology glossary. The test corpus contains representative English text entries from *Elden Ring*, especially samples with dense proper nouns, implicit background information, frequent attribute terms, and distinctive stylistic features. Each test sample is paired with a human reference translation for calculating BERTScore F1 and BLEU-4. The local semantic knowledge base contains background information related to world-building, character relationships, geographical regions, faction settings, equipment background, and terminology explanations. It is used only as an external contextual source during translation and is not involved in model training. The domain terminology glossary contains important character names, place names, equipment names, attribute terms, faction names, and world-building concepts. To handle possible translation variants, multiple acceptable translations are allowed for the same concept through alias matching.

For data processing, this study first cleans the test texts, reference translations, and model-generated translations by removing abnormal line breaks, carriage returns, redundant spaces, and irrelevant punctuation. The knowledge base texts are then organized by field and encoded with the MiniLM model. For each model-generated translation, semantic similarity, character-level BLEU-4, and terminology matching statistics are calculated. Since no new translation model is trained, the data are not divided into training, validation, and test sets in the traditional supervised-learning sense, but are organized as a test corpus, a semantic knowledge base, and a terminology evaluation glossary.

4.3. Evaluation metrics

To evaluate video game localization quality, this study adopts three dimensions: semantic similarity, surface-level lexical overlap, and domain terminology preservation. The metrics include BERTScore F1, character-level BLEU-4, terminology precision (Term-P), the Global Terminology Enrichment Score (GTES/Term-R), and terminology F1 (Term-F1). BERTScore F1 measures macro-level

DOI: 10.54254/2755-2721/2026.35240

semantic similarity, BLEU-4 measures surface-level overlap, and terminology-related metrics evaluate the preservation of game-domain terms.

4.3.1. BERTScore F1

BERTScore is an automatic evaluation metric based on contextual embeddings from pre-trained language models. It measures semantic similarity by matching textual units in the candidate translation and the reference translation. Its precision, recall, and F1 score are calculated as follows:

$$P_{BERT} = \frac{1}{|C|} \sum_{x_i \in C} \max_{y_j \in R} sim\left(x_i, y_j\right)$$

$$R_{BERT} = \frac{1}{|R|} \sum_{y_j \in R} \max_{x_i \in C} sim\left(x_i, y_j\right)$$

$$F1_{BERT} = 2 \times \frac{P_{BERT} \times R_{BERT}}{P_{BERT} + R_{BERT}} \tag{1}$$

where (C) and (R) denote the candidate translation and the reference translation, respectively; (x_i) and (y_j) denote their contextual embedding vectors; and (sim(x_i,y_j)) denotes semantic similarity between two vectors. In this study, BERTScore is calculated with lang="zh" and idf=False. The average BERTScore F1 across all samples is used as the final semantic similarity score for each experimental group.

4.3.2. Character-level BLEU-4

BLEU measures n-gram overlap between a candidate translation and a reference translation. This study uses character-level BLEU-4 to reduce the influence of Chinese word segmentation. The BLEU score is calculated as follows:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2}$$

where (N=4), (p_n) denotes modified n-gram precision, and (w_n) denotes the n-gram weight. In this study, the weights of 1-gram to 4-gram are equal, namely (w_n=0.25). (BP) denotes the brevity penalty:

DOI: 10.54254/2755-2721/2026.35240

$$BP = \begin{cases} 1, & c > r \\ e^{1-r/c}, & c \leq r \end{cases} \tag{3}$$

where (c) and (r) denote the lengths of the candidate translation and the reference translation. A smoothing function is applied to reduce the influence of missing higher-order n-grams in short texts, and the final BLEU-4 score is converted into a percentage.

4.3.3. Global Terminology Enrichment Score (GTES)

Since video game localization texts contain many proper nouns, place names, attribute terms, faction names, and world-building concepts, BERTScore F1 and BLEU-4 alone cannot fully reflect domain knowledge preservation. Therefore, this study introduces terminology precision, GTES, and terminology F1 to evaluate the model's ability to preserve game-domain terminology.

This study first constructs a domain terminology glossary with standard concept names, acceptable translation variants, and term weights. Let (T_{ref}) denote the term set in the reference translation, (T_{cand}) denote the term set in the candidate translation, and (w_t) denote the weight of term (t). Terminology precision is calculated as follows:

$$P_{term} = \frac{\sum_{t \in T_{ref} \cap T_{cand}} w_t}{\sum_{t \in T_{cand}} w_t} \tag{4}$$

GTES is defined as weighted terminology recall:

$$GTES = R_{term} = \frac{\sum_{t \in T_{ref} \cap T_{cand}} w_t}{\sum_{t \in T_{ref}} w_t} \tag{5}$$

Terminology F1 is calculated as follows:

$$F1_{term} = 2 \times \frac{P_{term} \times R_{term}}{P_{term} + R_{term}} \tag{6}$$

Here, (P_{term}) measures the accuracy of terminology usage, (R_{term}) or GTES measures the recall of reference terms, and (F1_{term}) balances terminology accuracy and completeness. This study uses GTES as the core domain-knowledge metric, while Term-P and Term-F1 serve as auxiliary indicators. For terminology-intensive and highly intertextual game texts such as those in *Elden Ring*, these metrics can more directly reflect whether context enhancement helps the model access and apply relevant background knowledge.

DOI: 10.54254/2755-2721/2026.35240

4.4. Analysis of ablation experiment results

The ablation experiment on GPT-5.5 was designed to separate the effects of prompt engineering and context enhancement. Group A uses only the source text, Group B adds the prompt engineering module, Group C adds retrieved context, and Group D combines both modules. The results are presented in Table 1.

Table 1. Ablation results on GPT-5.5

Experiment	PEM	CEM	BLEU-4	Term-P	GTES/Term-R	TermF1
GroupA			17.76	75.61	71.52	72.98
GroupB	√		16.14	72.52	67.07	67.39
GroupC		√	19.57	87.48	85.19	85.78
GroupD	√	√	21.99	86.62	94.12	89.19

The results show that prompt engineering alone does not improve GPT-5.5 in this setting. Compared with Group A, Group B decreases in BLEU-4, GTES, and Term-F1, which suggests that role descriptions and style instructions cannot by themselves supply the missing background knowledge in *Elden Ring* texts. The context enhancement group performs much better: GTES rises from 71.52% to 85.19%, and Term-F1 rises from 72.98% to 85.78%. When both modules are used together, Group D achieves the best overall result, with 21.99% in BLEU-4, 94.12% in GTES, and 89.19% in Term-F1. This pattern indicates that retrieved background information plays a more direct role than static prompting in terminology preservation, while the full method can further combine contextual support with style control.

To examine whether the method works across different models, this study also compares the full-method results of Doubao, Gemini, and GPT-5.5. The results are shown in Table 2.

Table 2. Cross-model results of the full method

Experiment	PEM	CEM	BLEU-4	Term-P	GTES/Term-R	TermF1
Db-GroupD	√	√	16.75	82.35	70.82	74.26
Gemini-GroupD	√	√	23.52	80.39	72.63	75.74
GPT-GroupD	√	√	21.99	86.62	94.12	89.19

The cross-model comparison shows that the full method improves domain terminology recall for all three models. In GTES, Doubao increases from 53.00% to 70.82%, Gemini from 58.13% to 72.63%, and GPT-5.5 from 71.52% to 94.12%. Term-F1 also rises from 54.94% to 74.26% for Doubao, from 61.67% to 75.74% for Gemini, and from 72.98% to 89.19% for GPT-5.5. Although the three models have different baseline capabilities, the same upward trend suggests that the proposed method has a certain degree of cross-model applicability.

Figure 2 shows the changes in BLEU-4, terminology precision, GTES, and Term-F1 across different experimental groups. Overall, the introduction of retrieved context leads to clearer improvement in terminology-related metrics than prompt engineering alone. The full method group performs most consistently in GTES and Term-F1, which further supports the role of context retrieval in preserving game-domain terminology and world-building references.

DOI: 10.54254/2755-2721/2026.35240

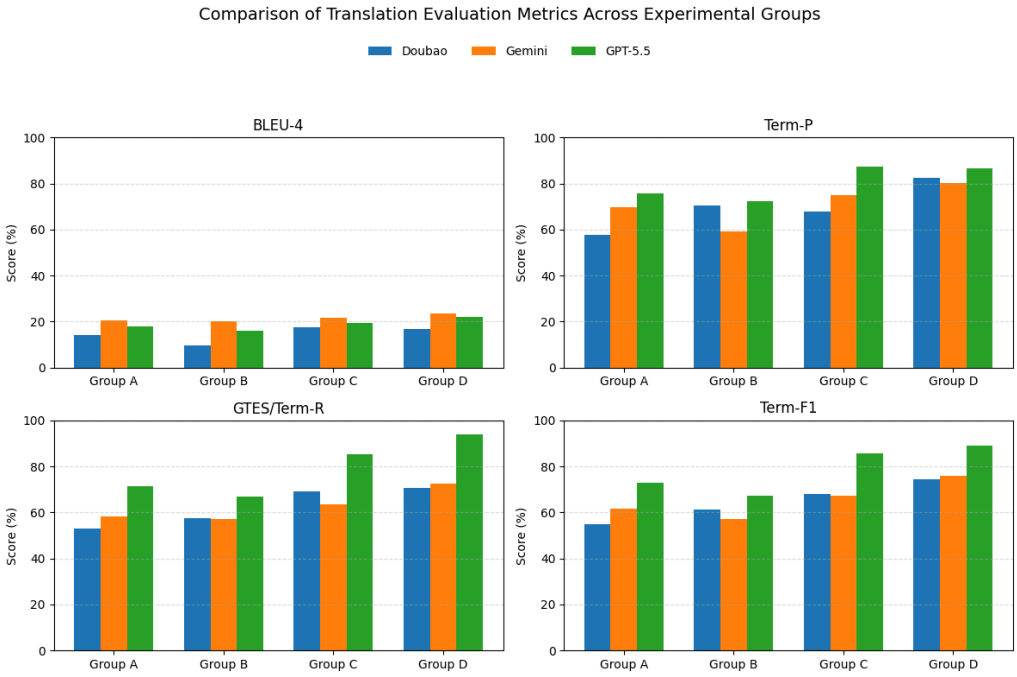

Figure 2. Metric changes across experimental groups

5. Conclusion and future work

This study started from a practical problem in game localization: many game texts are not translated as continuous narratives, but as separate strings. In a fragmented narrative game such as *Elden Ring*, an item description or a skill text may look short, yet it can be connected to character relations, faction history, place names, and recurring world-building terms. When such entries are translated without enough background information, large language models may still produce fluent Chinese, but the translation can lose important narrative links or use terms inconsistently.

To respond to this problem, this study designed a context-enhanced prompting framework. Instead of relying only on general translation prompts, the framework retrieves related background information from a local semantic knowledge base and adds it to the translation prompt. The experiment compared Doubao, Gemini, and GPT-5.5 under four settings: baseline translation, prompt engineering, context enhancement, and the full method. The results show that prompt engineering alone is not always reliable. In some cases, role and style instructions help control the tone, but they do not necessarily give the model the missing background knowledge. By contrast, context enhancement produces clearer improvement in terminology-related metrics. The full method achieves the best overall results, especially in GTES and Term-F1, which suggests that retrieved context can help the model preserve game-specific terms and world-building references more effectively.

The study also shows the value of GTES as a supplementary evaluation metric. BERTScore and BLEU-4 can reflect semantic similarity and surface overlap, but they are not designed specifically for terminology-heavy game texts. GTES focuses on whether important domain terms and world-building concepts are recalled in the translation, making it more suitable for evaluating fragmented game localization texts.

There are still several limitations. The corpus is mainly taken from one game, so the framework should be tested on more genres and language pairs in future work. The current knowledge base is

DOI: 10.54254/2755-2721/2026.35240

also text-based, while actual game localization may require visual, audio, and scene-level information. In addition, the terminology glossary and knowledge base are manually organized, which limits scalability. Future research can explore automatic terminology extraction, entity relation recognition, knowledge graph construction, Top-k retrieval, reranking, and context compression. Human evaluation should also be included, because automatic metrics cannot fully judge style, readability, cultural adaptation, or player acceptance. Overall, this study suggests that retrieval-enhanced prompting is a practical way to reduce contextual loss in fragmented game text translation and can serve as a useful reference for intelligent computer-assisted game localization.

References

[1] Jiao, W., Wang, W., Huang, J.T., Wang, X., Shi, S. and Tu, Z. (2023) Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine. arXiv preprint arXiv: 2301.08745.
[2] Yamada, M. (2023) Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability. In Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, 195-204.
[3] He, S. (2024) Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts. In Proceedings of the 25th Annual Conference of the European Association for Machine Translation, Volume 1, 316-326.
[4] He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S. and Wang, X. (2024) Exploring Human-Like Translation Strategy with Large Language Models. Transactions of the Association for Computational Linguistics, 12, 229-246.
[5] Miaschi, A. and Dell'Orletta, F. (2020) Contextual and Non-Contextual Word Embeddings: An In-Depth Linguistic Investigation. In Proceedings of the 5th Workshop on Representation Learning for NLP, 110-119.
[6] Chandler, H.M. and Deming, S.O. (2011) The Game Localization Handbook. 2nd Edition, Jones & Bartlett Publishers.
[7] O'Hagan, M. and Mangiron, C. (2013) Game Localization: Translating for the Global Digital Entertainment Industry. John Benjamins Publishing Company.
[8] O'Hagan, M. and Mangiron, C. (2006) Game Localisation: Unleashing Imagination with "Restricted" Translation. The Journal of Specialised Translation, 6, 10-21.
[9] Bernal-Merino, M.Á. (2006) On the Translation of Video Games. The Journal of Specialised Translation, 6, 22-36.
[10] Bernal-Merino, M.Á. (2015) Translation and Localisation in Video Games: Making Entertainment Software Global. Routledge.
[11] Hansen, D. and Houlmont, P.-Y. (2022) A Snapshot into the Possibility of Video Game Machine Translation. In Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Volume 2: Users and Providers Track and Government Track, 257-269.
[12] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S. and Kiela, D. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.
[13] Arslan, M., Ghanem, H., Munawar, S. and Cruz, C. (2024) A Survey on RAG with LLMs. Procedia Computer Science, 246, 3781-3790.